用户视角下农业科学数据描述信息的“结构–效用”研究

范智萱[1]，王　健[1]，撒　旭[1]，张贵兰[2]
（1.中国农业科学院农业信息研究所，北京　100081；2.中国科学技术信息研究所，北京　100038）

摘　要：[目的 / 意义]研究并提出科学数据描述信息的内容结构与其描述效用的对应关系，为科学数据描述的理论研究提供新的视角，为数字环境下农业科学数据的最优描述提供参考。[方法 / 过程]对 47 名农业领域硕博研究生被试的科学数据搜索与相关性判断行为进行准实验观察。首先，通过半结构化访谈获取被试相关性判断过程中使用的农业科学数据描述项集合及其使用特征；其次，分析高信心水平下用户的描述项使用路径；最后采用多元回归方法分析描述项对判断信心的预测能力。[结果 / 结论]研究得到了 11 类 42 项农业科学数据描述项，确定了来源、数据内容、使用与评价、数据产生信息是具备高效用的描述项，得到了高效用描述项组合，初步分析了用户数据素养和数据利用目的对描述项效用的影响。研究成果为科学数据元数据等具体的描述实践提供了理论依据。

关键词：科学数据；数据描述；元数据；信息效用；眼动追踪

中图分类号：G250　　文献标识码：A　　文章编号：1002-1248（2022）10-0057-13

引用本文：范智萱, 王健, 撒旭, 等. 用户视角下农业科学数据描述信息的“结构 - 效用”研究[J]. 农业图书情报学报, 2022, 34(10)：57-69.

1 引　言

在开放科学不断发展和社会经济数字化进程快速演进等因素的推动下，科学数据的资源总量和共享规模持续快速增长，已经成为了科技创新和社会经济发展的重要资源[1,2]。农业领域作为数据密集型的研究领域，横跨了生物学、生态学、作物学、气象学、食品工程等多个学科，其数据类型多样且规模庞大。近年来，国内外建立了多个农业科学数据库，农业科学数据的共享规模迅速增长且该领域人员的数据共享和重用意愿也在不断提升[3]。科学数据共享总量和重用需求的快速增长带来了数据发现和复用的挑战，传统的元数据和新兴的数据论文等数据描述形式不同程度地暴

收稿日期：2022-05-09

基金项目：国家农业科学数据中心平台运行经费项目“农业科学数据开放出版关键技术研究”（NASDC2022XM00-04）；国家科技基础条件平台中心委托课题“科学数据分级分类管理机制研究”

作者简介：范智萱（1996- ），硕士研究生，中国农业科学院农业信息研究所，研究方向为科学数据共享。王健（1971- ），博士，研究员，中国农业科学院农业信息研究所，研究方向为科学数据共享。撒旭（1997- ），硕士，中国农业科学院农业信息研究所，研究方向为科学数据共享。张贵兰（1993- ），博士，助理研究员，中国科学技术信息研究所，研究方向为科学数据共享

露出描述信息内容结构不充分、难以满足数据复用者信息需求的问题[4-6]，如何帮助复用者理解数据、判断数据的可复用性进而支持其实施数据复用行为，成为科学数据共享的核心议题[7-9]，也在实践层面引发了各类科学数据共享服务机构和科技期刊从不同角度开展的以扩增科学数据描述信息为主体的一系列描述增强行动[10]。

然而，此种做法及其所暴露的盲目性和低效果表明，扩增描述信息以增强描述效用（例如，支持数据复用者更准确地理解数据和做出更高质量的数据复用判断）的实践缺乏必要的理论指导和支持，做法背后的一系列科学问题——增加描述项（或调整描述信息的内容结构）是否必然导致其更好地满足数据复用者的信息需求（即描述信息效用的提升），哪些描述项具有更高的效用等[11]——都没有得到有效地探索和回答。显然，上述问题的核心可概括为描述信息的“结构 - 效用”关系。其中，“结构”是描述信息的内容组成，具体表现为元数据中的不同描述项组合或数据论文中的篇章结构；“效用”是科学数据描述信息对用户理解、判断进而施行复用行为的支撑程度；“结构 - 效用”关系反映了结构变化对效用水平的影响方向与程度。

论文将聚焦农业科学数据复用者的数据相关性判断过程，通过分析其描述信息的使用模式与特征探索不同描述项及其组合的效用变化，对描述信息的“结构 - 效用”关系开展初步探索，以期为农业科学数据描述实践发展提供参考和指导。

2 文献综述

2.1 科学数据描述

科学数据描述归源于信息描述，萌芽于世界数据中心早期“以物易物”式数据共享中的数据编目，其后在 E-Science 和开放科学（特别是开放数据）运动的接续推动下，逐渐发展形成了一个以“零描述”为起点，包含数据编目[12]、数据档案[13]、元数据[10]及增强元数据[14]、数据论文[15]等诸多中间形态，以理论上囊括全部描述信息的“全描述”为终点的描述谱系。该谱系清晰地勾画了科学数据描述随数据共享发展而呈现的信息不断丰富、形态不断多样的整体态势，也在一定程度上蕴含了描述信息的效用目标从数据治理向数据利用环节迁移扩充和效用水平不断提升的长期趋势。

科学数据元数据是目前最常用的描述方式，国内外已建立了数量众多的科学数据元数据标准体系。然而，一些过于复杂的标准增加了描述成本且难以应用于基于 Web 的检索系统中[16]。同时，现有标准也仍然存在描述效用不足的问题。例如与数据使用与评价相关的描述项受到用户的关注[17,18]且已被诸多学术检索系统整合到搜索结果展示中[19]，但却未包含在现有的元数据标准中。科学数据采集情境的描述信息与科学数据特性相关，但在很多标准中非必选项，甚至被忽略[20]。

数据论文是科学数据常见的描述方式之一，是用于展示大型数据集的一种独特的文章类型，包含了丰富的数据文档，对于数据重用而言至关重要。KIM[21]通过对 24 个数据期刊的数据论文指南进行内容分析，发现他们更多关注数据生产信息（数据收集、数据生产者和项目）和重用信息（潜在的重用和使用条款），而这些内容恰恰弥补了元数据的不足。数据论文的出现有效鼓励了个人或机构的数据共享，让科学数据发挥更多的潜在价值[15]。

2.2 科学数据描述信息使用研究

科学数据描述信息的使用研究能够发现用户判断科学数据的认知过程，并在此基础上评价了描述信息在用户判断时发挥的作用。对于实现科学数据的“有效描述”具有重要意义，也是目前一个重要的研究议题。

CHIN 和 LANSING[22]通过与生物学家们讨论数据共享和重用的不同场景确定了 11 类关键特征或属性，包括常规数据集属性、实验属性、数据来源、集合、分析和解释、物理组织、项目组织、科学组织、任务、实验过程和用户社区，形成了一个较为全面的信息框架。FANIEL 等[17]通过对社会学、考古学、动物学的研究人员进行研究，发现有关数据生产信息、存储库信息和数据使用信息是做出是否重用数据决策的关键。

KOESTEN 等[18]以创建用户为中心的数据摘要指南为目的，通过对 69 名学生的 269 份数据搜索日志进行编码，确定了涵盖评估数据集相关性、可用性和质量等不同方面的数据集属性列表。随后，该学者基于信息搜索行为模型，确定了用户理解数据过程中存在的检查、接触内容、将数据与不同情境相关联 3 种活动模式及其相关的数据属性[23]。

国内学者中，常颖聪等[24]通过对植物学领域 15 名博士生及研究人员进行访谈，并结合德尔菲专家调查法建立了植物学基因表达实验元数据模型，包括实验设计、实验数据、实验结果、科研成果、实验操作、数据访问和实验管理信息 7 个模块。赵华等[25]对 36 名农业领域的研究生开展了眼动实验和实验后访谈，研究指出发挥最大认知价值的元数据项依次是数据介绍（摘要）、数据来源、在线链接地址和关键词。除此之外，数据快照、同源数据、相关数据等非元数据项也发挥着重要作用。

通过文献调研发现，目前学者们针对科学数据描述信息的内容及结构开展了探索性和描述性研究，主要集中于识别一系列元数据元素，试图更加全面地描述数据集，但很少有研究关注所创建的元数据标准是否能够促进数据重用，即元数据元素的描述效用问题。因此，本研究将以科学数据为信息载体，通过情境实验、访谈、问卷调查、统计分析等定性定量相结合的方法，探究科学数据描述信息与其描述效用之间的关系，为更有效的科学数据描述提供参考。

3　理论框架及研究设计

3.1　理论框架

研究综合运用透镜模型、概率心理模型和适应性决策框架 3 个关键模型，建立了描述信息结构与描述效用之间关系的概念模型。透镜模型将人类判断过程中的要素分解为客观事物特征（效标，Criterion）、人类对这些特征的感知（线索，Cues）和由此形成的主观判断（Subject Judgments）3 个部分，构建了从感知信息到判断形成的认知过程[26]。概率心理模型（The Theory of Probabilistic Mental Models，PMM）与透镜模型理论密切相关，其基本观点是人们判断信心的形成与判断结果的产生既同时发生，也依靠同样线索。PPM 理论中的“线索有效性（Cue Validity）”概念与透镜模型中线索的“生态效度（Ecological Validity）”概念[27]具有较大相似性。适应性决策行为框架（Adaptive Decision Making）认为决策者的目标是最大限度地提高决策的准确性，同时最小化所投入的认知努力[28]。该理论解释了本研究探索的高效用描述项是用户权衡的结果，且描述信息的使用特征会受到任务情境的影响。

综合上述观点构建了本研究的理论模型（图 1）。透镜模型提供了整体框架，将描述信息感知与价值判断和信心达成等分为 4 个认知阶段。其中，描述项是科学数据描述信息的概念化，具体操作为用户在实验中接收的具有描述功能的语义视觉单元；中间是用户头脑中进行认知加工的过程，包含线索（即用户对描述项的感知）和标准（即用户赖以进行价值判断的个性化、工具性认知结构）两个要素；判断是用户对目标数据集相关性的感知，具体操作为用户对当前数据的相关程度的二值判断；判断信心反映用户对其判断结果的信心程度，是描述效用的概念化，在实验中通过李克特量表取得。

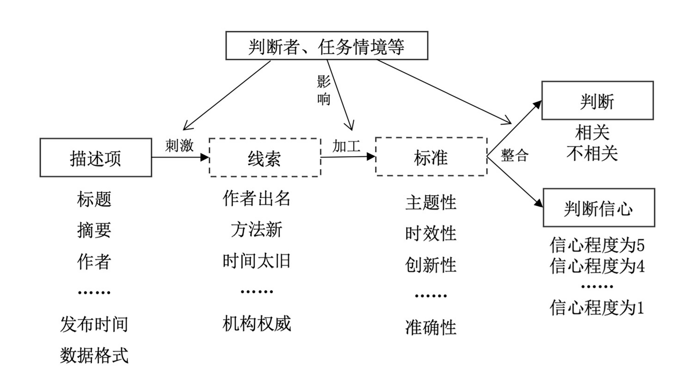


图 1　科学数据描述信息的“结构 - 效用”模型

Fig.1　Structure-utility model of descriptive information of scientific data

模型引出了当前研究力图回答的 3 个问题：①农业科学数据复用者使用哪些描述项及使用特征是什么？②不同描述项或其组合的描述效用如何？③影响描述

效用的因素有哪些以及如何影响？本研究将通过回答这 3 个问题建立描述信息结构与效用之间的定性关系。

3.2 实验设计

研究人员基于科学数据相关性判断场景构建了观察实验，具体包括 4 个步骤（图 2）。

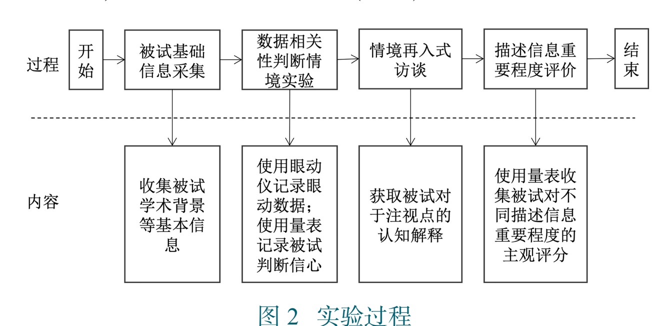


图 2 实验过程

Fig.2 Experimental process

（1）实验前被试基础信息采集。使用调查问卷采集被试 3 个方面的信息，包括：①专业背景、学历等个人基本信息；②科学数据检索频率、使用频率、检索平台等科学数据使用情况；③待检索主题、数据利用目的等检索相关信息。

（2）科学数据相关性判断及描述项使用观察。每名被试任选其熟悉的 1 个科学数据检索平台（中英文不限）作为操作环境，自选检索主题进行检索并逐一完成 10 次科学数据集相关性判断（相关 / 不相关），检索与判断均不限时。主试采用 Eye-Link2000 记录仪全程记录被试每次判断的眼动数据（注视点和眼动轨迹），同时采用 SR Research Screen Recorder1.0.0 录屏软件全程记录被试的操作。每次判断完成后被试需要通过 5 档量表（1= 完全没有信心，... ，5= 完全自信）给出其判断信心。

（3）相关性判断后的情境再入式访谈。每次判断和判断信心量表填写完成后，主试与被试共同观看被试的判断全过程录像与眼动轨迹，就关注点和眼动轨迹背后的认知加工进行访谈，访谈提纲详细表 1，并全程录音。

（4）描述信息重要程度评价。访谈结束后，被试需通过 5 档量表（1= 完全不重要，... ，5= 非常重要）给出不同描述信息的重要程度得分。

表 1 访谈提纲主要内容

Table 1 The main content of the interview outline

访谈提纲
1.你认为该数据集相关/不相关是基于什么信息判断的？
2.为什么要关注这个区域？这里给你带来了什么信息？
3.你认为这个数据集对你是否有用？具体用途是什么？
4.目前这些信息是否足够使你做出判断？还需要哪些信息？

研究人员通过社交平台微信，在中国农业科学院研究生院硕博各年级的学生群组中发布被试招募通知，遴选数据查询和使用经验较为丰富且来自多个学科的研究生。研究最终选取来自中国农业科学院 14 个研究所的 47 名研究生，其中硕士研究生 40 人，博士研究生 7 人；男性 15 人，女性 32 人。被试学科分布如表 2 所示。为激励学生参与实验，研究人员为每位被试发放了 100 元的津贴。

表 2 被试的学科分布

Table 2 Discipline distribution of subjects

所在学科	频数/次	百分比/%
食品科学与工程	12	25.5
农林经济管理	8	17.0
生物学	7	14.9
植物保护	7	14.9
园艺学	5	10.6
作物学	3	6.4
农业工程	1	2.1
大气科学	1	2.1
农业资源利用	1	2.1
畜牧学	1	2.1
测绘科学与技术	1	2.1
合计	47	100.0

3.3 数据收集和处理

3.3.1 访谈数据收集和处理

研究共收集到 47 份访谈录音，经转录后得到超过 16 万字的编码原始材料。研究人员依据 FANIEL 等[17]和 KOESTEN 等[18]的研究及科学数据元数据标准建立了初始编码表，使用 Nvivo11 软件进行内容分析并在编码过程中对初始编码表进行修改和扩充。两位编码

员分析了 43 份访谈文本并进行了交叉检验，同时预留了 4 份访谈文本作为饱和度校验样本。饱和度检验表明，编码结果均能纳入已形成的概念中，编码实例详见表 3。

3.3.2 问卷数据收集和处理

本实验收到 47 名被试的实验前、实验后问卷，所有问卷均在实验环境中填写，无重复答题的情况且被试的答题时间均在正常范围内，因此均为有效问卷。对问卷的处理主要包括两个部分：①使用实验前问卷收集的被试学历背景、数据获取能力（检索频率、常用的检索平台数量）、数据处理与分析能力（使用频率）3 类指标，并为每类指标赋予相同的权重，将被试的数据素养水平划分为高、中、低 3 个等级；根据数据利用目的，将被试划分为将数据作为参考和将数据加工应用两类，具体分布如表 4、表 5 所示。②从每位被试的情境实验中随机挑选 3 条数据集条目，共得到 140 条完整记录。结合编码结果确定被试在每条数据的判断过程中所关注的描述项，关注到的描述项被赋予实验后问卷中被试对描述项重要程度的打分，未关注的均为 1 分。

表 4 不同数据素养水平的被试分布

Table 4 Distribution of subjects with different levels of data literacy

数据素养水平	频数/次	百分比/%
高	13	27.6
中	24	51.1
低	10	21.3
合计	47	100.0

表 5 不同数据利用目的的被试分布

Table 5 Distribution of subjects with different data utilization purposes

数据利用目的	频数/次	百分比/%
作为参考	34	72.3
加工应用	13	27.7
合计	47	100.0

4 结 果

4.1 农业科学数据描述项集合及其使用频次

研究得到了 42 个描述项，并依据编码表将其归纳于 11 个类别，描述项的具体内涵与使用频次如表 6 所示。

4.1.1 高使用频次描述项

在 11 个类别中，用户使用频次最高的 5 个类别依次为主题（23.8%）、数据内容（23.7%）、整体描述（17.3%）、来源（13.5%）和数据产生信息（8.5%），合计为 86.8%，其余描述项的个体使用频次占总频次比例均不足 5%，合计频次占总频次比例不足 15%。在 42 个描述项中，使用频次排名前五的依次是标题、数据值、摘要、数据产生方法和实验结果。

（1）主题。反映目标科学数据集主题的描述项。所有 47 位用户均提及该类别中的描述项。用户在选择一个数据集时，首先需要判断数据的主题相关性[29]。当主题相关时，用户通常会继续寻找其他信息从而做出进一步判断；若主题无关，则会终止评估过程并做出放弃行为。正如用户所说“看完标题后觉得挺相关的，但需要再看更多的信息进一步确定”“看完题目可确定不是我想找的，所以没有再看摘要”。此外，由于同一研究主题在不同学科之间存在差异，因而用户也会进一步关注数据产生的学科领域。例如“该数据

表 3 编码实例

Table 3 Examples of coding process

访谈文本	描述项
用户 9：我觉得偏农业类的期刊的质量会好一点，像中国农业学报、中国农业气象比较权威	来源期刊
用户 27：因为收集数据的时间范围才到 07 年，这篇论文发布年份太老了，对我没用	发布时间、时间范围
用户 21：我觉得这篇更好的是它设置了 3 个不同浓度，比只有一个梯度的要好，做的比较深入	数据产生方法
用户 2：数据的值跟常识没有太大出入就觉得是合格的	数据值

表 6 农业科学数据描述项及其使用频次统计

Table 6 Descriptive items of agricultural scientific data and their use frequency statistics

描述项类别	定义	描述项	频次/次	百分比/%	合计/%
主题（Subject）	反映目标科学数据集的主题	标题	213	15.91	23.8
		主题概念	47	3.51	
		关键词	29	2.17	
		图表名称	22	1.64	
		学科领域	8	0.60	
数据内容（Data）	反映目标科学数据集的数据或图表本身	数据值	170	12.70	23.7
		实验结果	80	5.97	
		数据集表头	32	2.39	
		变量指标	28	2.09	
		图表横纵坐标	8	0.60	
整体描述（Overall Description）	反映目标科学数据集的具体内容	摘要	158	11.80	17.3
		数据集说明	73	5.45	
来源（Source）	反映目标科学数据集的来源期刊、机构或作者情况	来源期刊	68	5.08	13.5
		机构	36	2.69	
		数据来源	18	1.34	
		作者	18	1.34	
		影响因子	16	1.19	
		发布平台	11	0.82	
		基金资助	7	0.52	
		期刊类别	4	0.30	
		作者研究方向	3	0.22	
数据产生信息（Data Production Information）	反映目标科学数据集的产生过程和分析方法	数据产生方法	82	6.12	8.5
		实验材料	13	0.97	
		数据分析方法	13	0.97	
		研究思路	6	0.45	
范围（Coverage）	反映目标科学数据集覆盖的时间、空间范围	时间范围	37	2.76	4.1
		空间范围	18	1.34	
使用与评价（Use and Evaluation）	反映目标科学数据集的传播情况和用户评价	引用次数	36	2.69	3.7
		下载次数	8	0.60	
		获取限制	3	0.22	
		用户评价	2	0.15	
时间（Date）	反映目标科学数据集的时效性	发布时间	29	2.17	2.5
		数据更新频率	4	0.30	
物理特性（Physical Traits）	反映目标科学数据集的格式、类型、长度、大小	数据文件大小	13	0.97	1.3
		图表数量	3	0.22	
		数据格式	2	0.15	
质量（Quality）	反映目标科学数据集的质量	人工审核标记	5	0.37	0.8
		数据质量描述	3	0.22	
		数据缺失情况说明	3	0.22	
与其他信息的关联情况（Related Information）	反映目标科学数据集的关联信息	相似数据	4	0.30	0.7
		参考文献	4	0.30	
		相关数据库信息	2	0.15	

集里涉及很多化学方面的内容，但我不是研究化学的，所以不需要”。

（2）数据内容。反映目标科学数据集的数据或图表本身的描述项。共有 42 位用户提及该类别中的描述项。这些用户并非查看了数据集全集，而是仅关注了图表的表头、变量指标或某个结论性的数据值。表头是用户使用最普遍的数据属性之一，特别在结构化数据中会受到更多的关注[30]。例如“我一般先关注表头，如果表头相关，我会再看具体数值”。此外，结论性数据值被提及的频率远高于其他描述项，用户希望通过评估数据值是否符合内心预期的范围来评价数据的准确性。例如“一般看配比和功能系数，我有一个比较区间，功能系数比 120 大的越多，说明效果好”“我看氨基酸序列的长度，同源家族长度都相近，我会根据这个值判断是否再进一步检查”。

（3）整体描述。反映目标科学数据集具体内容的描述项。43 位用户提及该类别中的描述项，且几乎所有使用论文支撑性数据的用户都提到了摘要。尽管大多数数据集都具有标题、关键词等元数据，但这类简短词汇类型的元数据通常无法提供足够的内容让用户判断数据是否有用[31]，因此，文本类型的摘要或数据集说明则至关重要，能够提供更丰富的信息帮助用户评估数据的相关性、可用性和质量[32]。例如“这个数据是浏览了摘要后发现研究对象和研究方法都是非常相关的”“只看题目是相关的，但是摘要没有我想要的信息”。

（4）来源。反映目标科学数据集的来源期刊、机构或作者情况的描述项。32 位用户提及该类别中的描述项。与数据集来源有关的信息有助于用户评估数据集的质量、权威性或可信度[17]。例如用户提到“我关注了机构，这个学校在这方面的研究位于领域前沿，老师也很有名”“根据期刊名称可以判断期刊的好坏，也就能判断是否要下载”“我认为国家统计局的数据相对权威一些”。

（5）数据产生信息。反映目标科学数据集的产生过程和分析方法的描述项。39 名用户提及该类别中的描述项。文献中的数据提供了更多的与数据产生相关的信息，包括整体的研究思路、收集数据的实验设置等。当用户认为数据生产者所使用的方法与自己的实验方法类似时，往往会对该数据集的相关程度给予较高评价。例如“实验方法、还有抗菌肽、分子设计等，跟我的研究相关性比较高”。同样，该类别还有助于用户评估数据集的新颖性，相较于陈旧的方法而言，使用更加新颖、创新的方法所产生的数据通常会得到用户较高的评价。“他的方法比较创新，而且它的检测线比较低，得出的规律是挺好的”。

此外，用户还会对数据产生过程的规范性、实验设置的合理性做出评价，当用户对数据产生过程存在质疑时，则会降低其评价。例如“这是实际生产中去做的，跟我的不相关，并且我认为这种方式不太正规”。

4.1.2　其他描述项

除上文中介绍的 5 类描述项外，其余 6 个类别在部分用户的判断过程中也发挥了作用。

（1）范围。反映目标科学数据集覆盖的时间、空间范围的描述项。24 位用户提及该类别中的描述项。空间范围包括收集数据的地点、层级（例如国家、省、市等），尤其当用户的研究与地理环境相关时，会对该描述项更感兴趣。“它是一个县域层面的考虑，我想要的全国类的，所以我选择放弃”。时间范围能够让用户筛选出他们不感兴趣的数据集，如果时间覆盖范围不是他们所需的，用户则很容易做出放弃的决定。但在更多情况下，当用户发现数据仅覆盖了他们所需的部分时期时，仍然会选择下载该数据集并之后再自行补充。例如“数据不新肯定有影响，但是我部分可以参考它，我的数据集会涵盖很多年份，它是我的一个子集”。

（2）使用与评价。反映目标科学数据集的传播情况和用户评价的描述项。26 位用户提及该类别中的描述项。与数据使用情况相关的信息反映了同一学科群体中人们对该数据集的认可和接受水平。其中，提及频次最高的描述项是引用次数，一般在用户评价数据集的质量时出现。例如“引用次数也能说明它的质量、权威性”。当然，这类描述项在用户判断中并不会起到决定性作用，用户还会综合考虑数据集的发布时间、

学科领域特性等因素。

此外，部分用户指出他人对于数据的评价能提供更多的信息，尤其是当自己的判断还不太确定时。“如果数据有问题，会有网友评论数据不能使用或者存在错误，这些评论会影响我对它的判断”。评价信息能够帮助用户提前评估目标信息的有用性和质量[33]，使用户在付出较少认知努力的情况下做出信心充分的判断[34]。

（3）时间。反映目标科学数据集的时效性的描述项。20 位用户提及该类别中的描述项。同一个数据集在不同的时间下具有不同的价值。在很多研究主题下，用户都倾向于获取最新的数据，而较为陈旧的数据则会被放弃。例如“很早之前的数据就不太会考虑，一般都是要最新的”“我没有找到我想要的，这个数据没有及时更新”。

（4）物理特性。反映目标科学数据集的格式、类型、大小的描述项。7 位用户提及该类别中的描述项，它们不表达数据集的主题或内容，而是侧重于数据集的物理特征。在本研究中，用户大多通过数据格式判断数据集是否可用，通过数据大小评估数据的质量和缺失情况。正如用户提到“这个是时间序列的，是我想要序列性的数据”“文件大小能够让我判断数据下载下来是不是我想要的，如果它很小就说明数据可能会存在缺失”。

（5）质量。反映目标科学数据集质量的描述项。7 位用户提及该类别中的描述项，且均为在数据共享平台中检索数据时提及。这种情况的出现与平台本身有关，一方面是数据共享平台中可能会包含数据质量说明文档，而文献平台中没有；另一方面文献出版已经过同行评审，因而不会再带有人工审核标志，但在数据共享平台中这些信息都是对数据集质量最直接的说明。“质量描述会介绍数据的缺失情况和数据质量文档的查看位置，我会关注数据缺失情况，但需要看具体的质量文档才能判断的”“这个数据集没有出现人工审核标志，我觉得没有必要点进去看”。

（6）与其他信息的关联情况。反映目标科学数据集关联信息的描述项。5 位用户提及该类别中的描述项。部分用户使用关联信息来辅助评估数据的质量，当他们认为参考文献的数量丰富或质量较高时，会提升对整体研究质量的判断。例如用户在访谈中提到“参考文献的数量能够表现出他一部分质量”“当与某个蛋白质相关的数据库数量较多时，说明更多人对其进行了研究，数据的可信度则更高”。

4.2 高效用描述项组合分析

通过分析用户在科学数据相关性判断过程中描述项的使用，发现绝大多数情况下用户需依赖两个及以上的描述项做出决策。因此，本节将通过分析用户的描述项使用路径，识别具备高效用（支持用户做出高信心水平的判断）的描述项组合。本文选取用户信心充分的判断过程（判断信心为 4 分及以上）进行描述项使用路径分析，发现所有用户均会首先关注主题或整体描述类别中的属性，以判断当前数据的主题相关性。因此，当用户除关注主题或整体描述外还使用了其他描述信息时，主题和整体描述被归为一类进行统计，以消除不同路径之间的内涵重复。

研究人员对判断结果为“相关”的访谈记录进行编码，共得到 25 条使用路径，其中合计频次占比约 80%的前 8 种使用路径如表 7 所示。其中，使用频次最高的前 3 种描述项组合依次为主题或整体描述、数据产生信息、数据内容；主题或整体描述、数据内容；主题或整体描述、来源、数据产生信息、数据内容。值得注意的是，68.75%的用户在检查完数据内容相关描述项后即可结束判断过程，由此可推断数据内容能够提供给用户更直观、更具说服力的信息，对于用户判断信心有显著影响。

研究人员对判断结果为“不相关”的访谈记录进行编码，共得到 16 条使用路径，其中合计频次占比约 80%的前 6 种使用路径如表 8 所示。与做出一个“相关”的判断相比，用户做出一个“不相关”的判断往往只需要较少的信息即可达到较高的信心水平，且有 22.5%的用户只需要依据主题或主题和整体描述信息即可决定放弃该数据集。此外，没有用户关注到使用与评价信息和与其他信息的关联情况两个类别，这两类

表 7　高信心水平下的描述项使用路径（判断结果为“相关”时）

Table 7 Descriptive items usage path at high confidence level (when the result is "relevant")

描述项使用路径	频数/次	百分比/%
主题（整体描述）→数据产生信息→数据内容	32	20.9
主题（整体描述）→数据内容	25	16.3
主题（整体描述）→来源→数据产生信息→数据内容	21	13.7
主题（整体描述）→来源	16	10.5
主题→整体描述	9	5.9
主题（整体描述）→来源→数据内容	7	4.6
主题（整体描述）→范围→来源→数据内容	6	3.9
主题（整体描述）→时间→来源→数据产生信息→数据内容	6	3.9
合计	122	79.7

表 8　高信心水平下的描述项使用路径（判断结果为“不相关”时）

Table 8 Descriptive items usage path at high confidence level (when the result is "irrelevant")

描述项使用路径	频数/次	百分比/%
主题（整体描述）→数据产生信息	17	23.6
主题（整体描述）→数据内容	13	18.1
主题→整体描述	10	13.9
主题（整体描述）→数据产生信息→数据内容	7	9.7
主题	6	8.3
主题（整体描述）→来源	5	6.9
合计	58	80.6

信息只在判断结果为“相关”的场景下发挥作用。

4.3　高效用描述项及其对判断信心的影响

4.3.1　描述项对判断信心的整体影响

对描述项与用户判断信心的回归分析表明 11 类描述项对判断信心变化影响显著，R^2=0.31，p<0.01。如表 9 所示，在 11 个类别中，来源、数据内容、使用与评价、数据产生信息被认为是用户判断信心的显著预测因素，且均为正相关（在 α=0.05 的水平上）。主题和整体描述类别在判断过程中的普遍使用使其无法在统计学层面体现出对判断信心的显著预测性。来源、数据内容、数据产生信息 3 类描述信息对于判断信心的正向影响解释了它们对于科学数据用户的必要性。使用与评价信息的使用频率较低，但在面临特定用户和特定数据利用目的的情况下对判断信心具有显著的正向影响。

4.3.2　不同数据素养水平下描述项对判断信心的影响

在影响用户判断过程中的因素中，判断者被诸多学者认为是影响最大的因素[35]，包含判断者的学科背景[36]、专业水平[17]、判断能力[37]等。对于科学数据而言，科学数据素养这一概念很好地概括了科学数据用户收集、加工、管理、评价和利用数据的能力与知识[38]。本文使用多元回归分析检验不同数据素养水平的用户所依赖的描述信息的差异性。结果说明，对于科学数据素养较低的用户，影响其判断信心的重要因素是数据来源（R^2=0.28，p<0.01）；对于中等水平的用户，影响其判断信心的重要因素是使用与评价和数据产生信息（R^2=0.27，p<0.05）；对于高水平的用户，数据的使用和评价仍然很重要，但用户还会更倾向于依赖数据内容（R^2=0.31，p<0.01）。

4.3.3　不同数据利用目的下描述项对判断信心的影响

在本研究中，因用户的专业性质和研究方向不同，

表 9　描述项重要程度与判断信心回归分析

Table 9 Regression analysis of the importance of descriptive items and judgment confidence

描述信息结构	总体样本	按数据素养水平分类			按数据利用目的分类	
		低	中	高	参考	加工
主题						
整体描述						
来源	0.164*	0.384**				
范围						
时间						
物理特性						
数据内容	0.139*			0.275*		0.591*
使用与评价	0.180*		0.217*	0.257*	0.188*	
与其他信息的关联情况						
数据产生信息	0.207***		0.176*		0.212**	
质量						

* 注：* 表示 $P < 0.05$，** 表示 $P < 0.01$，*** 表示 $P < 0.001$

其检索科学数据的目的或意图也有所差异。将检索到的数据作为参考（例如作为研究背景、数据对比或数据补充）的用户，影响其判断信心的重要因素是使用与评价和数据产生信息（R^2=0.26，p<0.01），表明他们更关注于数据产生过程和被认可度。将检索到的数据进行加工应用的用户，则更倾向于依赖数据本身（R^2=0.41，p<0.01），在本实验中，该类用户一般通过专业的数据共享平台获取数据，例如国家统计局、气象局、NCBI 等，因而很少会质疑数据的权威性或产生过程，更多关注的是数据本身是否符合要求以及数据是否可用。

5　结论与建议

良好的描述是科学数据高效治理、传播、发现和利用的必要前提。不同的描述信息组合表现了不同的描述效用，进而演化形成了从“零描述”到“充分描述”的科学数据描述连续统。描述信息的“结构 - 效用”关系是连续统演化的理论基础和内在逻辑，也是应对当前描述能力不足问题的良好进路。本研究以实证方式识别了 11 类 42 项农业科学数据描述项及其使用特征，确定了高效用描述项与描述项组合，并确定了用户数据素养和数据利用目的是影响描述信息效用的两个关键变量及其影响。

本研究根据用户使用频次，得到发挥主要作用的农业科学数据描述项包括主题、数据内容、整体描述、来源和数据产生信息 5 个类别。与农业科学数据核心元数据标准[39]进行对比发现，在现有标准中，对于主题、整体描述和来源信息的描述较为充分，但数据产生信息仅作为数据质量模块中的一个非必选项出现，数据内容类别中的描述项则未有体现。两者之间的差异为农业科学数据元数据的完善提供了一些方向。例如，可考虑增设描述数据集内容的元数据模块，包含表头、图表横纵坐标、结论性数值等易于提取且能够以标准化形式呈现的描述项。特别是对于结构化的数据集而言，创建该模块可使用户对其形成更加直观的了解。此外，还应加大数据产生信息的描述力度，可为其单独设置一个元数据模块并提供更加细化的元数据元素。

对于科学数据共享平台，建议将用户使用率高的描述项优先呈现并确保其完整性和准确性，将具备高效用的描述项放置在明显位置并考虑更优化的呈现方式。例如可提供诸如数据快照、关键图表的缩略图等直观反映数据集内容的描述信息、提供数据评价或数

据评分、特定时间范围内的下载量或引用量，例如近一周内、近一个月内、近一年内等。此外，可将用户关注度较低的描述项进行折叠以突出关键信息，有助于建立用户友好型的数据共享平台。

本研究在一定意义上为传统上开展的、实践主导的元数据研究和数据论文研究提供了一个认知导向的整合框架，可为数字环境下科学数据元数据和新型描述形式发展提供理论参考和指导。但本文仍存在一定局限性，研究的样本群体均是来自于农业领域的硕博研究生，在其他学科领域和其他用户群体中缺乏代表性。其次，本研究识别出的用户所关注和使用的描述项是对实验环境的反映和适应。当我们提供给用户其他的描述信息时，结果可能会发生一些改变。因此，未来这项研究需要在更大规模的不同人群中反复进行，且应在不同的任务情境中进一步调查。

参考文献：

[1] YOON A. Data reusers' trust development[J]. Journal of the association for information science and technology, 2017, 68(4): 946-956.

[2] 孙玉伟, 成颖, 谢娟. 科研人员数据复用行为研究: 系统综述与元综合[J]. 中国图书馆学报, 2019, 45(3): 110-130.
SUN Y W, CHENG Y, XIE J. A review on the data reuse behavior of scholars: System review and meta synthesis[J]. Journal of library science in China, 2019, 45(3): 110-130.

[3] TENOPIR C, ALLARD S, DOUGLASS K, et al. Data sharing by scientists: Practices and perceptions[J]. Plos one, 2011, 6(6): 1-21.

[4] FEAR K. User understanding of metadata in digital image collections: Or, what exactly do you mean by" coverage"?[J]. The American archivist, 2010, 73(1): 26-60.

[5] NIU J. Perceived documentation quality of social science data[M]. University of Michigan, 2009.

[6] TENOPIR C, ALLARD S, DOUGLASS K, et al. Data sharing by scientists: Practices and perceptions[J]. PloS one, 2011, 6(6): E21101.

[7] BIRNHOLTZ J P, BIETZ M J. Data at work: Supporting sharing in science and engineering[C]//SCHMIDT K, PENDERGAS M, TREMAINE M, SIMONE C (eds). Proceedings of the 2003 international ACM SIGGROUP conference on supporting group work, Sanibel Island, Florida, USA: ACM, 2003: 339-348.

[8] BORGMAN C L. Big data, little data, no data: Scholarship in the networked world[M]. MIT press, 2017.

[9] GREGORY K, GROTH P, SCHARNHORST A, et al. Lost or found? Discovering data needed for research[J]. Harvard data science review, 2020, 2(2): 1-63.

[10] 邱春艳. 国内外科学数据相关实践中的元数据研究进展[J]. 情报资料工作, 2021, 42(5): 104-112.
QIU C Y. Research progress of metadata in scientific data-related practices at home and abroad [J]. Information and documentation services, 2021, 42(5): 104-112.

[11] DREWRY M, CONOVER H, MCCOY S, et al. Metadata: Quality vs. quantity[C]. IEEE computer society metadata conference, 1997: 1-11.

[12] 苏广利, 姜翠景. 机读目录格式与元数据格式用于网络资源组织的比较研究[J]. 图书馆杂志, 2002(1): 19-22.
SU G L, JIANG C J. Research on application of metadata and marc in organizing web resources[J]. Library journal, 2002(1): 19-22.

[13] 花文博. 浅论基础地理信息数据档案的管理[J]. 兰台世界, 2010(s2): 128-129.
HUA W B. Talking about the management of basic geographic information data archives[J]. Lantai world, 2010(s2): 128-129.

[14] CHAO T C. Enhancing metadata for research methods in data curation[J]. Proceedings of the American society for information science and technology, 2014, 51(1): 1-4.

[15] 刘凤红, 崔金钟, 韩芳桥, 等. 数据论文: 大数据时代新兴学术论文出版类型探讨[J]. 中国科技期刊研究, 2014, 25(12): 1451-1456.
LIU F H, CUI J Z, HAN F Q, et al. Data papers: Discussion on the emerging academic paper publishing types in the era of big data[J]. Chinese journal of scientific and technical periodicals, 2014, 25 (12): 1451-1456.

[16] RIALL R L, MARINCIONI F, LIGHTSOM F L. Content metadata standards for marine science: A case study[M]. US department of the interior, US geological survey, 2004.

[17] FANIEL I M, FRANK R D, YAKEL E. Context from the data reuser's point of view[J]. Journal of documentation, 2019, 75(6): 1274-1297.

[18] KOESTEN L, SIMPERL E, EMILIA M K, et al. Everything you always wanted to know about a dataset: Studies in data summarisation [J]. International journal of human-computer studies, 2020, 135: 1-21.

[19] BEHNERT C. Investigating the effects of popularity data on predictive relevance judgments in academic search systems[C]. Proceedings of the 2019 conference on human information interaction and retrieval, 2019: 437-440.

[20] 满芮, 王健. 我国农业元数据标准初探[J]. 中国科技资源导刊, 2016, 48(2): 7-12.

MAN R, WANG J. Analysis of status quo about metadata standard in the agricultural field[J]. China science & technology resources review, 2016, 48(2): 7-12.

[21] KIM J. An analysis of data paper templates and guidelines: Types of contextual information described by data journals [J]. Science editing, 2020, 7(1): 16-23.

[22] CHIN JR G, LANSING C S. Capturing and supporting contexts for scientific data sharing via the biological sciences collaboratory[C]. Proceedings of the 2004 ACM conference on computer supported cooperative work, 2004: 409-418.

[23] KOESTEN L, GREGORY K, GROTH P, et al. Talking datasets-understanding data sensemaking behaviours[J]. International journal of human-computer studies, 2021, 146: 102562.

[24] 常颖聪, 何琳. 科学实验数据元数据模型构建研究——以植物学基因表达实验为例[J]. 图书情报工作, 2015, 59(13): 117-125.

CHANG Y C, HE L. Research on construction of metadata model for scientific experimental data: An example as gene expression experiment of botany[J]. Library and information service, 2015, 59 (13): 117-125.

[25] 赵华, 周国民, 王健, 等. 科学数据元数据认知价值评价研究[J]. 情报科学, 2016, 34(7): 81-85.

ZHAO H, ZHOU G M, WANG J, et al. Study of evaluation of cognitive value of scientific metadata[J]. Information science, 2016, 34 (7): 81-85.

[26] 海斯蒂, 道斯. 不确定世界的理性选择[M]. 北京: 人民邮电出版社, 2013.

HASTIE R, DAWES R M. Rational choice in an uncertain world[M]. Beijing: Posts & telecom press, 2013.

[27] GIGERENZER G, HOFFRAGE U, KLEINB?LTING H. Probabilistic mental models: A brunswikian theory of confidence[J]. Psychological review, 1991, 98(4): 506.

[28] PAYNE J W, PAYNE J W, BETTMAN J R, et al. The adaptive decision maker[M]. Cambridge university press, 1993.

[29] 刘建平. 科学数据用户相关性判断模型研究[D]. 北京: 中国农业科学院, 2020.

LIU J P. Research on relevance judgement model of scientific data users[D]. Beijing: Chinese academy of agricultural sciences, 2020.

[30] KOESTEN L. A user centered perspective on structured data discovery[C]. Companion proceedings of the web conference 2018, 2018: 849-853.

[31] BRICKLEY D, BURGESS M, NOY N. Google dataset search: Building a search engine for datasets in an open web ecosystem[C]. The world wide web conference, 2019: 1365-1375.

[32] AU V, THOMAS P, JAYASINGHE G K. Query-biased summaries for tabular data[C]. Proceedings of the 21st Australasian document computing symposium, 2016: 69-72.

[33] YOU S. Evaluative metadata in educational digital libraries: How users use evaluative metadata in the process of document selection[J]. TCDL bulletin, 2010, 4(2): 1-11.

[34] TANG M C. A study of academic library users' decision-making process: A lens model approach[J]. Journal of documentation, 2009, 65(6): 938-957.

[35] 于春, 彭爱东, 王波, 等. 信息用户对信息检索相关性判断的因素分析[J]. 图书情报工作, 2009, 53(3): 103-107.

YU C, PENG A D, WANG B, et al. The relevance factors of judgment in the information retrieval of the information users[J]. Library and information service, 2009, 53(3): 103-107.

[36] GREGORY K, GROTH P, COUSIJN H, et al. Searching data: A review of observational data retrieval practices in selected disciplines [J]. Journal of the association for information science and technology, 2019, 70(5): 419-432.

[37] ROUET J F, ROS C, GOUMI A, et al. The influence of surface and deep cues on primary and secondary school students' assessment of relevance in web menus[J]. Learning and instruction, 2011, 21(2): 205-219.

[38] 孟祥保, 李爱国. 国外高校图书馆科学数据素养教育研究[J]. 大学图书馆学报, 2014, 32(3): 11–16.
MENG X B, LI A G. Scientific data literacy education in overseas academic libraries[J]. Journal of academic libraries, 2014, 32(3): 11–16.

[39] 孟宪学. 农业科学数据核心元数据 I[M]. 北京: 中国农业科学技术出版社, 2008.
MENG X X. Agricultural scientific data core metadata I[M]. Beijing: China agriculture science and technology press, 2008.

Structure-Utility of Descriptive Information of Agricultural Scientific Data from the Perspective of Users

FAN Zhixuan[1], WANG Jian[1], SA Xu[1], ZHANG Guilan[2]
(1. Agricultural Information Institute, Chinese Academy of Agricultural Sciences, Beijing 10008;
2. Institute of Scientific and Technical Information of China, Beijing 100038)

Abstract: [Purpose/Significance] This paper aims to study the structure-utility relationship of descriptive information of scientific data to provide a new perspective for the theoretical study of scientific data description and a reference for the best description of agricultural scientific data in the digital environment. [Method/Process] Based on information processing theory, the lens model, the probabilistic mental model theory and the adaptive decision-making behavior framework, the relationship model between descriptive information structure and informing utility was constructed. A situational experiment was designed according to the model. In this study, 47 postgraduates from 14 institutes were invited for quasi-experimental observation by using qualitative and quantitative methods such as eye-tracking, semi-structured interview and questionnaire. First, this study used a semi-structured interview to obtain a user's cognitive interpretation of fixation points and collected the descriptive items of agricultural scientific data and their use frequency by encoding the interview text. Second, this study combined descriptive item usage path coding and user judgment confidence to obtain the combination of descriptive items with high utility. Finally, the study used multiple regression analysis to identify the descriptive items with high utility and their predictive ability, and analyzed the impact of data literacy and data utilization type on the utility of descriptive items. [Results/Conclusions] The study identified 42 descriptive items of 11 categories of agricultural scientific data and their usage characteristics. Among them, the top 5 frequently used descriptive items were subject, data, overall description, source and data production information, which played an important role in user relevance judgment. Then this study identified the combination of descriptive items with high utility and found that users' use patterns of descriptive items were diverse. Compared with making a judgment with "relevant" result, users often needed less information to achieve a high level of confidence when making an "irrelevant" judgment. This study also found that the descriptive items with high utility include source, data, use and evaluation, and data production information. It is determined that user data literacy and data utilization purpose were the influencing factors of descriptive information utility, and the effects of the two factors were preliminarily analyzed. Based on this research, the paper put forward some suggestions for improving agricultural scientific data metadata and scientific data sharing. In the future, this study will be repeated in groups with different academic backgrounds and data literacy levels, so as to enhance the generalization ability of research conclusions and construct a more effective structure of scientific data descriptive information.

Keywords: scientific data; data description; metadata; information utility; eye-tracking